

COSPEDTree-II: Improved Couplet based Phylogenetic Supertree

Sourya Bhattacharyya, and Jayanta Mukhopadhyay

Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur,
West Bengal 721302, India
sourya.bhatta@gmail.com, jay@cse.iitkgp.ernet.in

Abstract—A Supertree synthesizes the topologies of a set of phylogenetic trees carrying overlapping taxa set. In process, conflicts in the tree topologies are aimed to be resolved with the consensus clades. Such a problem is proved to be NP-hard. Various heuristics on local search, maximum parsimony, graph cut, etc. lead to different supertree approaches, of which the most popular methods are based on analyzing fixed size subtree topologies (such as triplets or quartets). Time and space complexities of these methods, however, depend on the subtree size considered. Our earlier proposed supertree method COSPEDTree, uses evolutionary relationship among individual couplets (taxa pair), to produce slightly conservative (not fully resolved) supertrees. Here we propose its improved version COSPEDTree-II, which produces better resolved supertree with lower number of missing branches, and incurs much lower running time. Results on biological datasets show that COSPEDTree-II belongs to the category of high performance and computationally efficient supertree methods.

Index Terms—Phylogenetic tree, Supertree, Couplet, Directed Acyclic Graph (DAG), Equivalence Relation, Transitive reduction, Internode count.

I. INTRODUCTION

Supertree methods combine the evolutionary relationships of a set of phylogenetic trees \mathbf{G} , into a single tree \mathbf{T} [1]. These methods differ from the consensus-based approaches [2], [3], by allowing input trees to have different but overlapping set of taxa. Supertrees are useful in combining input trees generated from completely incomparable approaches, such as statistical analysis of discrete dataset and distance analysis of DNA-DNA hybridization data [1]. Input trees often exhibit conflicting topologies, due to different evolutionary histories of respective genes, stochastic errors in site and taxon sampling, and biological errors due to paralogy, incomplete lineage sorting, or horizontal gene transfer [4]. Supertree methods quest for resolving such conflicts in order to produce a ‘median tree’, which minimizes the sum of a given distance measure with respect to the input trees [4]. Large scale supertrees are intended towards assembling the *Tree of Life* [5].

Our earlier work [6], and the study in [1], provide a comprehensive review of various supertree methods. *Indirect* supertree methods first generate intermediate structures like matrices (as in MRP [7], Minflip [8], SFIT [9]) or graphs (as in Mincut (MC) [10], modified Mincut [11]) from the input trees, and subsequently resolve these intermediate structures to produce the final supertree. These methods, especially MRP, are quite accurate, but computationally intensive. *Direct*

methods, on the other hand, derives the supertree directly from input tree topologies. These methods may aim for minimizing either the sum of false positive (FP) branches (as in the *veto* approaches like PhySIC [12], SCM [13]) or the sum of Robinson-Foulds (RF) [14] distance (as in RFS [15]) between \mathbf{T} and \mathbf{G} . Another approach named Superfine [16], [17] employs greedy heuristics on MRP [7] or Quartet Maxcut (QMC) [18], to derive the supertree, which may not be completely resolved. Supertrees formed by synthesizing the subtrees (such as triplets [4], [19], quartets [18], [5]) of the input trees, exhibit quite high performance. But, time and space complexities of these methods depend on the size of the subtree used.

We have previously developed COSPEDTree [6], a supertree algorithm using evolutionary relationships among individual pair of taxa (couplets). The method is computationally efficient, but produces somewhat conservative (not fully resolved) supertrees, with low number of false positive (FP) but high number of false negative (FN) branches between \mathbf{T} and \mathbf{G} . Here we propose its improved version, termed as COSPEDTree-II, which produces better resolved supertree, with lower number of FN branches between \mathbf{T} and \mathbf{G} , keeping the FP count also low. We have also proposed a mechanism to convert a non-resolved supertree into a strict binary tree, to reduce the FN count. COSPEDTree-II requires significantly lower running time than COSPEDTree and most of the reference methods, particularly for the datasets having high number of trees or taxa.

Rest of this manuscript is organized as follows. First, we review the basics of COSPEDTree (as in [6]) in section II. The method COSPEDTree-II is then described in section III. Performance of COSPEDTree-II is summarized in section IV.

II. OVERVIEW OF COSPEDTREE

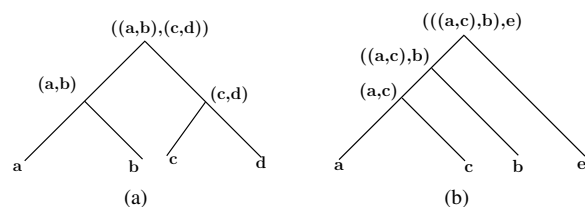


Fig. 1. Example input phylogenetic trees. All the nodes are labeled by Newick [20] representation.

Let \mathbf{G} consist of M rooted input trees t_1, t_2, \dots, t_M . For an input tree t_j ($1 \leq j \leq M$), let $L(t_j)$ be its set of constituent taxa. Suppose a pair of taxa p and q belong to $L(t_j)$. Further, let ϕ_p and ϕ_q be the parent internal nodes (points of speciation) of p and q , respectively. COSPEDTree [6] defines four boolean relations r_k^{pq} ($k \in \{1, 2, 3, 4\}$) between p and q , with respect to t_j , as:

- 1) Earlier Speciation of p than q (r_1^{pq}) is true, if ϕ_p is ancestor of ϕ_q in t_j . For the tree in Fig. 1(b), r_1^{bX} is true, where $X \in \{a, c\}$. Similarly, r_1^{cY} is true for $Y \in \{a, c, b\}$.
- 2) Later Speciation of p than q (r_2^{pq}) is true, if ϕ_p is a descendant of ϕ_q . So, r_1^{qp} and r_2^{pq} are equivalent.
- 3) Simultaneous Speciation of p and q (r_3^{pq}) is true, if $\phi_p = \phi_q$. In Fig. 1(a), r_3^{ab} and r_3^{cd} are true.
- 4) Incomparable Speciation of p and q (r_4^{pq}) is true, when ϕ_p and ϕ_q occur at different (and independent) clades. For the tree in Fig. 1(a), r_4^{ac} is true.

Using another taxon $s \in L(t_j)$, properties of r_1 to r_4 can be stated as the following:

P1: Both r_1 and r_2 are transitive. Thus,

- $r_1^{pq} \& r_1^{qs} \Rightarrow r_1^{ps}$.
- $r_2^{pq} \& r_2^{qs} \Rightarrow r_2^{ps}$.

P2: r_3 is an equivalence relation.

P3: $r_3^{pq} (= r_3^{qp}) \& r_k^{ps} \Rightarrow r_k^{qs}$, where $k \in \{1, 2, 4\}$.

P4: $r_1^{pq} (= r_2^{qp}) \& r_4^{ps} (= r_4^{sp}) \Rightarrow r_4^{qs} (= r_4^{sq})$.

Support tree set Γ_{pq} for a couplet (p, q) is defined as:

$$\Gamma_{pq} = \{t_j : (p, q) \in L(t_j)\} \quad (1)$$

The frequency F_k^{pq} ($k \in \{1, 2, 3, 4\}$) of a relation r_k^{pq} between a couplet (p, q) is the number of input trees t_j where $t_j \in \Gamma_{pq}$ and r_k^{pq} is true.

The set of allowed relations $R(p, q)$ between a couplet (p, q) is defined as the following:

$$R(p, q) = \{r_k^{pq} : F_k^{pq} > 0\} \quad (2)$$

A couplet (p, q) exhibits *conflict* if $|R(p, q)| \geq 2$ (where $|\cdot|$ denotes the cardinality of a set). The *consensus relation* between (p, q) is the relation having the maximum frequency.

Priority measure P_k^{pq} for a relation r_k^{pq} ($k \in \{1, 2, 3, 4\}$) between a couplet (p, q) is defined as the following:

$$P_k^{pq} = F_k^{pq} - \sum_{1 \leq k' \leq 4, k' \neq k} F_{k'}^{pq} \quad (3)$$

COSPEDTree also defines a *support score* V_k^{pq} for individual relations r_k^{pq} as the following:

$$V_k^{pq} = F_k^{pq} \times P_k^{pq} \quad (4)$$

The consensus relation r_k^{pq} between a couplet (p, q) exhibits the highest P_k^{pq} and F_k^{pq} values. So, corresponding V_k^{pq} also becomes the highest among all relations between (p, q) .

Final supertree \mathbf{T} resolves (assigns a particular relation to) individual couplet (p, q) with a single relation r_k^{pq} ($k \in \{1, 2, 3, 4\}$) between them. Maximum agreement property

[15] of a supertree quests for resolving individual couplets with their respective consensus relations. But, satisfying such property is NP-hard since consensus relations among couplets can be mutually conflicting [6]. Thus, order of selection of individual candidate relations r_k^{pq} (to resolve the corresponding couplet (p, q)) is crucial. In view of this, COSPEDTree first constructs a set of relations S_r , such that if a relation $r_k^{pq} \in S_r$, the couplet (p, q) is resolved with r_k^{pq} . To construct S_r , COSPEDTree applies an iterative greedy approach. At each iteration, it selects a relation $r_k^{p'q'}$ to resolve (p', q') among all unresolved couplets, provided:

$$V_{k'}^{p'q'} = \max_{\forall (p, q), \forall k} V_k^{pq}.$$

If the selected relation $r_k^{p'q'}$ does not contradict with any of the already selected relations in S_r (according to the properties P1 to P4 mentioned before), it is included in S_r .

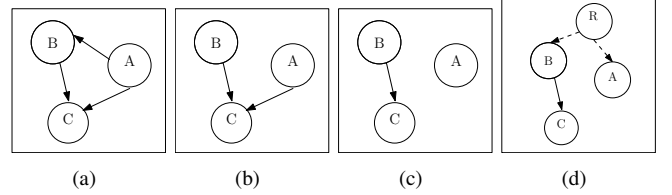


Fig. 2. (a) Transitive parent problem (b) Multiple parent problem MPP. (c) No parent problem NPP. (d) solution of NPP by COSPEDTree, by inserting a hypothetical root R.

Suppose, $L(\mathbf{G}) = \cup_{j=1}^M L(t_j)$ denotes the complete set of input taxa. Then, $N = |L(\mathbf{G})|$. Using the set of relations S_r , COSPEDTree partitions $L(\mathbf{G})$ into s mutually exclusive taxa clusters C_1, C_2, \dots, C_s , with the following rule (details are provided in [6]):

- R1: If a pair of taxa p and q belong to the same cluster C_i ($1 \leq i \leq s$), $r_3^{pq} \in S_r$.
- R2: Suppose C_i and C_j ($1 \leq i, j \leq s$, $i \neq j$) are two distinct taxa clusters. Then, $\forall p \in C_i$, and $\forall q \in C_j$, $r_k^{pq} \in S_r$, where $k \in \{1, 2, 4\}$. This property is denoted by saying that $r_k^{C_i C_j}$ is true, or C_i is related with C_j via the relation r_k .

COSPEDTree creates a directed acyclic graph (DAG), whose nodes are individual taxa clusters C_i ($1 \leq i \leq s$). A directed edge from C_i to C_j means $r_1^{C_i C_j}$ is true. However, occurrence of one or more of the following properties means this DAG needs to be refined to form a tree:

- 1) Transitive parent problem (TPP): for three nodes A, B, and C, when r_1^{AC} , r_1^{BC} , and r_1^{AB} are simultaneously true, as indicated in Fig. 2(a).
- 2) Multiple parent problem (MPP): when r_1^{AC} , r_1^{BC} , and r_4^{AB} are simultaneously true, as shown in Fig. 2(b).
- 3) No parent problem (NPP) (Fig. 2(c)): when a node A does not have any parent, i.e. So, there exists no node B such that r_1^{BA} is true.

COSPEDTree [6] applies transitive reduction to resolve TPP. The problem MPP is solved by arbitrary parent assignment, while NPP is resolved by assigning one hypothetical

root node to the isolated node (as shown in Fig. 2(d)). Finally, a depth first traversal of this DAG produces the supertree \mathbf{T} . As there is no restriction regarding the number of taxa in individual taxa clusters (partitions with respect to the relation r_3), \mathbf{T} may not be strictly binary (completely resolved).

III. PROPOSED METHODOLOGY

COSPEDTree-II extends COSPEDTree by incorporating the following modifications:

1) COSPEDTree-II skips the formation of S_r . Rather, the taxa clusters (containing one or more taxon) are first derived, solely by the frequencies of different relations between individual couplets. Subsequently, directed edges between individual pairs of clusters are assigned, according to the properties of individual couplets contained within these cluster pairs. Such processing on the taxa clusters, rather than the couplets, achieves high speedup and much lower running time.

2) In COSPEDTree, if a relation r_k^{pq} ($1 \leq k \leq 4$) between a couplet (p, q) is supported in a tree $t_j \in \Gamma_{pq}$, the frequency F_k^{pq} is incremented by 1. COSPEDTree-II, on the other hand, uses fractional and dynamic frequency values. In the above case, COSPEDTree-II increments F_k^{pq} with a *weight* $W_{pq}^{t_j}$ ($0 < W_{pq}^{t_j} \leq 1$), which varies for individual couplets (p, q) , and also for individual trees $t_j \in \Gamma_{pq}$.

3) For the problem MPP, COSPEDTree-II proposes a deterministic selection of the parent, for the node having multiple parents.

4) COSPEDTree-II also suggests a mechanism to convert a non-binary supertree into a binary tree.

Subsequent sections describe all such improvements.

A. Fractional frequency value for relations

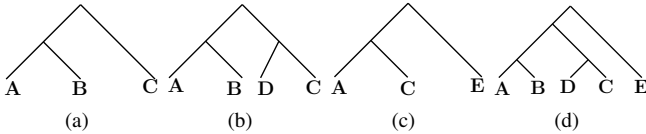


Fig. 3. Fig. (a) to (c) shows three input trees. Fig. (d) shows the corresponding supertree.

COSPEDTree-II applies a fractional frequency value $W_{pq}^{t_j}$ if an input tree t_j supports the relation r_k^{pq} between a couplet (p, q) . Value of $W_{pq}^{t_j}$ depends on the set $L(t_j)$. Utility of such a dynamic (and fractional) frequency measure is explained by Fig. 3, which shows three input trees (Fig. 3(a) to Fig. 3(c)) and corresponding supertree (Fig. 3(d)). For the couplet (A,C), all of the relations r_2 , r_3 and r_4 are supported. However, we observe that the relation r_3 is supported only because corresponding tree does not include taxa B and D. Similarly, the relation r_2 occurs due to the absence of the taxon D. When both B and D are present (Fig. 3(c)), the relation r_4 (which is the ideal relation between (A,C)) is satisfied. So, the relation r_4 should be given higher weight, since the corresponding tree has higher taxa coverage. So, our proposed dynamic frequency measure varies according to the coverage of taxa of different input trees.

Considering an input tree t_j ($1 \leq j \leq M$) and a couplet (p, q) in $L(t_j)$, first we define the following notations:

- $V(t_j)$: set of nodes (leaf or internal) of t_j .
- $LCA_{pq}^{t_j}$: lowest common ancestor (LCA) of p and q in t_j .
- $Clade_{t_j}(v)$: subtree rooted at an internal node $v \in (V(t_j) - L(t_j))$.
- $Cluster_{t_j}(v)$: Set of taxa underlying $Clade_{t_j}(v)$.

With such definitions, the set of *excess taxa* (excluding the couplet itself) underlying the LCA node of (p, q) in t_j , is defined as the following:

$$U_{pq}^{t_j} = Cluster_{t_j}(LCA_{pq}^{t_j}) - \{p, q\} \quad (5)$$

For (p, q) , union of all *excess taxa* underlying the respective $LCA_{pq}^{t_j}$ nodes for all $t_j \in \Gamma_{pq}$, is:

$$U_{pq}^{\mathbf{G}} = \bigcup_{t_j \in \Gamma_{pq}} U_{pq}^{t_j} \quad (6)$$

We assign the weight of a relation r_k^{pq} ($1 \leq k \leq 4$) between (p, q) in an input tree t_j , as:

$$W_{pq}^{t_j} = \frac{|U_{pq}^{\mathbf{G}} \cap L(t_j)|}{|U_{pq}^{\mathbf{G}}|} \quad (7)$$

where $W_{pq}^{t_j} = 1$ if $U_{pq}^{\mathbf{G}} = \phi$.

Thus, the weight equals the proportion of taxa within $U_{pq}^{\mathbf{G}}$, that is covered in the input tree t_j . Frequency F_k^{pq} of the relation r_k^{pq} , is now redefined as the following:

$$F_k^{pq} = \sum_{t_j \text{ supports } r_k^{pq}} W_{pq}^{t_j} \quad (8)$$

B. Generating taxa clusters

COSPEDTree [6] creates taxa clusters after formation of the *set of resolving relations* S_r . COSPEDTree-II, on the other hand, creates taxa clusters before resolving any couplets at all. Rather, for individual couplets (p, q) , COSPEDTree-II inspects the values of F_k^{pq} for individual relations r_k^{pq} ($k \in \{1, 2, 3, 4\}$). Creation of taxa clusters requires identifying couplets which can be resolved by the relation r_3 . COSPEDTree-II places a pair of taxa p and q in the same taxa cluster (thereby resolving the couplet (p, q) with the relation r_3), provided:

- 1) Either $|R(p, q)| = 1$ and $r_3^{pq} \in R(p, q)$ ($R(p, q)$ is already defined in Eq. 2).
- 2) Or $|R(p, q)| = 2$ and r_3^{pq} is majority consensus. In such a case, $F_3^{pq} \geq 0.5 * \{\sum_k F_k^{pq}\}$.
- 3) If $|R(p, q)| > 2$, the couplet (p, q) is not placed in the same taxa cluster, even if r_3^{pq} is majority consensus. This is because, as the couplet exhibits high degree of conflict, we check the relations between p , q , and other taxa set.

The first condition is obvious. A couplet having only r_3 as its allowed relation would be preferably resolved with it. On the other hand, if there exists one more relation $r_{k'}^{pq}$ ($k' \neq 3$)

within $R(p, q)$, we check whether $F_3^{pq} > F_{k'}^{pq}$, which ensures that r_3^{pq} is the majority consensus relation of (p, q) . In such a case, the couplet is highly probable of being resolved with r_3 in the final supertree.

Above mentioned heuristics are applied for individual couplets, to perform the equivalence partitioning (taxa clusters) of the input taxa set $L(\mathbf{G})$.

C. Connectivity between taxa clusters to form DAG

Creation of the taxa clusters is followed by the assignment of directed edges between them. As mentioned in section II, directed edge from a cluster C_i to a cluster C_j corresponds to the relation $r_1^{C_i C_j}$ ($= r_2^{C_j C_i}$) being true. In such a case, the cluster pair (C_i, C_j) is said to be *resolved by the relation* r_1 . In general, a pair of clusters can be resolved via any one of the relations r_1 , r_2 or r_4 (no directed edge in this case). For individual relations r_k ($k \in \{1, 2, 4\}$), we define its frequency $F_k^{C_i C_j}$ with respect to the pair of cluster (C_i, C_j) , as the following:

$$F_k^{C_i C_j} = \sum_{\forall p \in C_i, \forall q \in C_j} F_k^{pq} \quad (9)$$

Priority of individual relations r_k ($k \in \{1, 2, 4\}$) for the cluster pair (C_i, C_j) is defined as the following:

$$P_k^{C_i C_j} = F_k^{C_i C_j} - \sum_{k' \in \{1, 2, 4\}, k \neq k'} F_{k'}^{C_i C_j} \quad (10)$$

Support score of a relation r_k between the cluster pair (C_i, C_j) is defined as:

$$V_k^{C_i C_j} = P_k^{C_i C_j} + F_k^{C_i C_j} \quad (11)$$

Note that we have used sum, rather than the product, of the priority and frequency measures. This is due to the disparity of signs of frequency (which is always non-negative) and the priority (which can be negative even for a consensus relation) measures. Higher support score of a relation (between a pair of clusters) indicates higher frequency and priority of the corresponding relation.

The set Q of support scores for different relations between individual cluster pairs is defined as follows:

$$Q = \{V_k^{C_i C_j} : C_i \neq C_j, k \in \{1, 2, 4\}, F_k^{C_i C_j} > 0\} \quad (12)$$

Individual taxa clusters are now resolved by an iterative algorithm, using the set Q . Each iteration extracts a relation $r_{k'}^{C_x C_y}$ ($k' \in \{1, 2, 4\}$) from Q , provided the following:

$$V_{k'}^{C_x C_y} = \max_{\forall (C_i, C_j), \forall k} V_k^{C_i C_j} \quad (13)$$

Following conditions are checked to see whether the extracted relation $r_{k'}$ can resolve the cluster pair (C_x, C_y) .

- 1) If (C_x, C_y) is already resolved with a different relation, $r_{k'}$ is not applied.
- 2) If $k' = 1$ or 2 , resolving (C_x, C_y) with $r_{k'}$ would create a directed edge between the cluster pair. If such an edge forms a cycle with the existing configuration of the taxa clusters, $r_{k'}$ is not applied.

For no such above mentioned conflicts, the relation $r_{k'}$ is applied between C_x and C_y .

The set Q is implemented as a max-priority queue [21], to achieve $O(1)$ time complexity for extracting the cluster pair having the maximum support score. Iterations continue until Q becomes empty. However, the final DAG may still have the problems TPP, MPP, and NPP (as defined in Fig. 2). The problem TPP is removed by transitive reduction (already described in COSPEDTree [6]). COSPEDTree-II employs a better solution for the problem MPP, which is described in the following section.

D. Solving Multiple Parent Problem (MPP)

As shown in Fig. 2(b), the problem MPP corresponds to a cluster C_z having k ($k \geq 2$) other clusters C_1, C_2, \dots, C_k as its parent, which are not themselves connected by any directed edges. The objective is to assign a unique parent C_p ($1 \leq p \leq k$) to the cluster C_z . Such assignment was arbitrary in COSPEDTree [6]. COSPEDTree-II proposes a deterministic selection of C_p , by a measure called the *internode count* $I_{t_j}(p, q)$ between a couplet (p, q) , with respect to a rooted tree t_j . The measure was introduced in [22] for unrooted trees. Here, the measure is adapted for a rooted tree t_j , as the number of internal nodes between p and q through the node $LCA_{pq}^{t_j}$.

As individual trees t_j carry overlapping taxa subsets of $L(\mathbf{G})$, we define a *normalized internode count distance* between p and q in t_j as:

$$I_{t_j}^N(p, q) = \frac{I_{t_j}(p, q)}{W_{pq}^{t_j}} \quad (14)$$

where $W_{pq}^{t_j}$ is defined in the Eq. 7. So, $I_{t_j}^N(p, q)$ becomes low only when both $I_{t_j}(p, q)$ is low and $W_{pq}^{t_j}$ is high (when the tree t_j carries higher proportion of the taxa subset belonging to $U_{pq}^{\mathbf{G}}$).

Significance of the internode count distance can be explained by considering a rooted triplet $(r, (p, q))$ (shown in the Newick [20] format), consisting of three taxa p, q and r . Here, $I^N(p, q) < I^N(p, r) = I^N(q, r)$. In general, lower internode count means corresponding couplet is evolutionarily closer, compared to the other couplets.

Average internode count of a couplet (p, q) , with respect to \mathbf{G} , is defined by the following expression:

$$I_{avg}(p, q) = \frac{1}{|I_{pq}|} \sum_{t_j \in I_{pq}} I_{t_j}^N(p, q) \quad (15)$$

The *internode count distance between a pair of cluster C_x and C_y* is defined by the following equation:

$$I(C_x, C_y) = \frac{\sum_{\forall p \in C_x, q \in C_y} I_{avg}(p, q)}{|C_x||C_y|} \quad (16)$$

where $|C_x|$ denotes the cardinality of the taxa cluster C_x .

For the MPP problem, COSPEDTree-II selects the cluster C_p ($1 \leq p \leq k$) as the parent of C_z , provided that C_p has the lowest internode count distance to C_z :

$$C_p = \operatorname{argmin}_{1 \leq i \leq k} I(C_z, C_i) \quad (17)$$

Such condition is based on the assumption that the cluster pair having lower internode count, is possibly closer in the evolutionary tree, compared to other cluster pairs.

E. Binary supertree generation

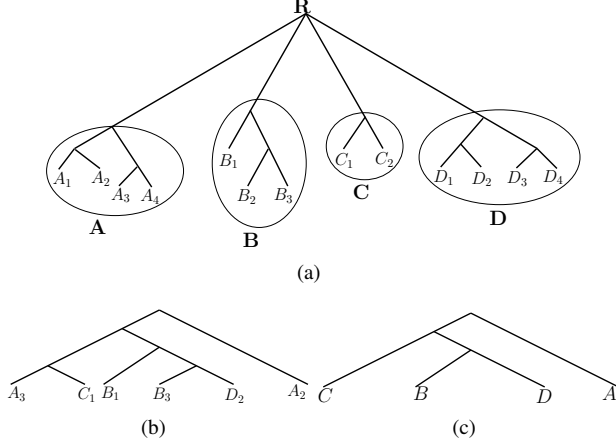


Fig. 4. (a) Example of a multifurcation, containing the taxa subset $X = A \cup B \cup C \cup D$. (b) an input tree $t_{j|X}$, restricted to the taxa subset X . (c) Tree $t'_{j|X}$ created from $t_{j|X}$.

After resolving the problem MPP, the refined DAG is converted to the supertree \mathbf{T} , by a depth first traversal procedure (as described in COSPEDTree [6]). However, the generated supertree \mathbf{T} may not be completely resolved. COSPEDTree-II proposes a refinement strategy which converts \mathbf{T} into a strict binary tree.

Suppose, the tree contains an internal multi-furcating node of degree n (> 2). Let X_1, X_2, \dots, X_n denote the taxa subsets descendant from it, where each taxa subset X_i ($1 \leq i \leq n$) consists of one or more taxon named as X_{i1}, X_{i2}, \dots , etc. Union of these taxa subsets is represented by $X = \cup_{i=1}^n X_i$. Suppose, $t_{j|X}$ represents the input tree t_j ($1 \leq j \leq M$) restricted to the set of taxa X . Thus, $L(t_{j|X}) = L(t_j) \cap X$. Considering Fig. 4(a) as an example, the node R represents a multi-furcation with degree 4. Four taxa subsets A, B, C, and D, descend from R . Here, $X = A \cup B \cup C \cup D$. Generation of a binary tree requires introducing bifurcations among these taxa subsets. So, for individual input trees t_j , corresponding restricted input tree $t_{j|X}$ is produced, as shown in Fig. 4(b).

Our proposed binary refinement first generates a tree $t'_{j|X}$ from the tree $t_{j|X}$, such that the leaves of $t'_{j|X}$ represent individual taxa subsets X_i ($1 \leq i \leq n$). In other words, individual taxon in $t_{j|X}$ is replaced by the name of its corresponding taxa subset (without any duplicate). For example, both the taxa A_2 and A_3 (belonging to the taxa subset A) are present in the tree $t_{j|X}$ (as shown in Fig. 4(b)). So, in $t'_{j|X}$, a leaf node labeled A is first inserted as a child of the LCA node of A_2 and A_3 . Subsequently, the leaves A_2 and A_3 are deleted from $t'_{j|X}$. This process is repeated for other taxa subsets B, C and D as well. Fig. 4(c) shows the tree $t'_{j|X}$.

For the current set of taxa X , each of the input trees t_j are processed to generate the corresponding $t'_{j|X}$. These trees

TABLE I
RESULTS FOR MARSUPIALS [23] DATASET ($M = 158, N = 267$)

Method	FP	FN	RF	MAST	Runtime
Minflip* [8]	792	946	1738	75.84	20m
MMC* [11]	911	1251	2162	69.3	-
MRP PAUP* [7]	756	400	1156	86.59	4.6m
PhySIC [12]	0	1324	1324	35.1	14s
RFS [15]	710	361	1071	105.6	6.5m
SCM* [13]	0	1220	1220	40.75	-
SFIT* [9]	1327	979	2306	61.69	111h
Superfine(MRP)* [16]	750	396	1346	89.54	3m
thSPR [19]	740	860	1600	97.26	3m
thTBR [19]	739	859	1598	97.1	3m
Supertriplet [4]	598	390	988	ER	57s
COSPEDTree	326	841	1167	68.93	6m
COSPEDTree-II	459	695	1154	78.02	2.4s + 2.2s
COSPEDTree-II+B	827	482	1309	89.34	2.4s+2.2s+34.4s

TABLE II
RESULTS FOR PLACENTAL MAMMALS [24] ($M = 726, N = 116$)

Method	FP	FN	RF	MAST	Runtime
Minflip* [8]	2965	4002	6967	276.64	7.25m
MRP PAUP* [7]	2545	1902	4447	313.77	3.5m
PhySIC [12]	0	4830	4830	222.86	3s
RFS [15]	2481	1650	4131	511.54	4.2m
SCM* [13]	0	4816	4816	223.86	-
SFIT* [9]	3315	2353	5668	316.53	108h
QIMP* [25]	2423	1823	4246	477.84	-
Superfine(MRP)* [16]	2540	1746	4286	439.65	9.25m
Superfine(QMC)* [16]	2631	1835	4466	432.84	6m
thSPR [19]	2627	3268	5895	496.23	7s
thTBR [19]	2626	3272	5898	496.25	6.39s
Supertriplet [4]	F	F	F	F	F
COSPEDTree	1232	3762	4994	394.42	3.2m
COSPEDTree-II	1406	2601	4007	429.34	3.5s + 0.5s
COSPEDTree-II+B	2730	1819	4549	491.85	3.5s+0.5s+9.2s

are then used as input to an existing triplet based supertree approach thTBR [19], to generate a supertree T_X consisting of the taxa subsets X_i as its leaves. The supertree method is selected since it processes rooted triplets, and generates a rooted output tree. The tree T_X is used as a template, such that its order of bifurcation among individual taxa subsets X_i is replicated to the original multi-furcating node R and its descendants. As the degree of multifurcation (n in this case) is much lower than the total number of taxa (N), construction of T_X is very fast. This process is continued until all the multi-furcating nodes are resolved.

TABLE III
RESULTS FOR SEABIRDS [26] DATASET ($M = 7, N = 121$)

Method	FP	FN	RF	MAST	Runtime
Minflip* [8]	38	72	110	3.24	11s
MRP PAUP* [7]	61	166	227	4.97	11s
PhySIC [12]	0	150	150	0.64	3s
RFS [15]	28	14	42	5.63	4s
SCM* [13]	1	66	67	2.75	-
SFIT* [9]	42	202	244	2.22	1h
Superfine(MRP)* [16]	32	19	51	4.43	3s
Superfine(QMC)* [16]	29	19	48	4.6	3s
thSPR [19]	69	242	311	5.71	6s
thTBR [19]	115	234	349	5.73	6s
Supertriplet [4]	2	176	178	ER	5.6s
COSPEDTree	24	81	105	2.37	3s
COSPEDTree-II	27	66	93	2.58	0.2s + 0.3s
COSPEDTree-II+B	50	37	86	5.02	0.2s+0.3s+3.1s

TABLE IV
RESULTS FOR THPL [27] DATASET ($M = 19$, $N = 558$)

Method	FP	FN	RF	MAST	Runtime
Minflip* [8]	142	149	291	4.93	1.1h
MRP PAUP* [7]	75	476	551	6.27	31m
PhySIC [12]	0	279	279	1.19	5.7m
RFS [15]	106	66	172	11.9	4.5m
SCM* [13]	13	128	141	4.64	-
Superfine(MRP)* [16]	85	50	135	6.39	1m
Superfine(QMC)* [16]	62	43	105	6.5	1.5m
thSPR, thTBR [19]	ER	ER	ER	ER	ER
Supertriplet [4]	F	F	F	F	F
COSPEDTree	88	162	250	4.21	4.5m
COSPEDTree-II	96	137	233	5.74	2s + 9s
COSPEDTree-II+B	166	114	280	8.22	2s+9s+1.2m

TABLE V
RESULTS FOR CETARTIODACTYLA [28] DATASET ($M = 201$, $N = 299$)

Method	FP	FN	RF	MAST	Runtime
MMC* [11]	1181	1438	2619	83.84	-
MRP PAUP* [7]	860	964	1824	120.84	-
PhySIC [12]	ER	ER	ER	ER	ER
RFS [15]	ER	ER	ER	ER	ER
thSPR [19]	969	1006	1975	118.39	5.5m
thTBR [19]	969	1006	1975	118.09	4.5m
Supertriplet [4]	125	2175	2300	ER	59s
COSPEDTree	510	1001	1511	80.43	11.7m
COSPEDTree-II	732	864	1566	95.2	2s + 1s
COSPEDTree-II+B	1240	667	1907	102.18	2s+1s+43s

F. Computational complexity of COSPEDTree-II

For M input trees covering a total of N taxa, both COSPEDTree [6] and COSPEDTree-II incurs $O(MN^2)$ time complexity for extracting the couplet based measures from the trees. These methods differ in their subsequent steps. COSPEDTree first resolves individual couplets in $O(N^2 \lg N)$ time (as shown in [6]), and subsequently partitions the taxa set according to the relation r_3 , to form a DAG containing N_C ($< N$) nodes (taxa clusters). Formation of a supertree from this DAG involves $O(N_C^3)$ time complexity [6].

COSPEDTree-II, on the other hand, first forms the taxa clusters in $O(N^2)$ time (processing time for all couplets). Subsequently, support scores for individual relations between each pair of taxa clusters are placed in the max-priority queue Q . Here, size of Q is $O(N_C^2)$, considering N_C as the number of taxa clusters. During each iteration, maintaining the max-priority property of Q requires $O(\lg N_C)$ time. So, the complete iterative stage to resolve all pairs of clusters (assigning connectivities between them) involves $O(N_C^2 \lg N_C)$ time complexity. As in general, N_C is considerably lower than N , this iterative step in COSPEDTree-II is much faster than COSPEDTree.

TABLE VI
RESULTS FOR MAMMAL [12], [4] DATASET ($M = 12958$, $N = 33$)

Method	FP	FN	RF	MAST	Runtime
PhySIC [12]	17414	254527	271941	968	36s
RFS [15]	ER	ER	ER	ER	ER
thSPR [19]	272752	296159	568911	8378	7s
thTBR [19]	276104	301787	577891	8378	7s
Supertriplet [4]	71117	105671	176788	ER	6s
COSPEDTree	35124	141295	176419	4441.5	5.3m
COSPEDTree-II	39079	134834	173913	4577.26	2m+0.1s
COSPEDTree-II+B	104226	98407	202633	8365.57	2m+0.1s+1m

Resolving individual pair of clusters, rather than the couplets, enables COSPEDTree-II to achieve a significant speedup. Suppose, $|X|$ denotes the cardinality of a taxa cluster X . So, for a pair of taxa clusters X and Y , COSPEDTree [6] resolves all $|X| \times |Y|$ couplets, and maintains their relations (and the transitive connectivities inferred from these relations). But COSPEDTree-II resolves X and Y by processing only one relation between them. So, for this cluster pair, speedup achieved by COSPEDTree-II is $\approx |X| \times |Y|$. For a total of N_C taxa clusters, number of cluster pairs is $\binom{N_C}{2}$. Thus, overall speedup achieved by COSPEDTree-II is $\approx \sum_{X,Y \in \binom{N_C}{2}}$

$|X| \times |Y|$.

To derive the time complexity associated with the binary refinement of COSPEDTree-II, suppose m is the number of internal nodes in T having degree > 2 . Further, suppose n (> 2) denotes the maximum degree of multi-furcation among all of these nodes. In such a case, applying thTBR [19] for a particular internal node involves maximum $O(Mn^3)$ time complexity. So, overall complexity of the binary refinement stage is $O(Mn^3m)$.

COSPEDTree [6] involves a storage complexity of $O(N^2)$, to store the couplet based measures. COSPEDTree-II uses additional storage space for storing the set of excess taxa U_{pq}^G for individual couplets (p, q) . As $0 \leq |U_{pq}^G| \leq (N - 2)$, the space complexity of COSPEDTree-II is $O(N^3)$.

IV. EXPERIMENTAL RESULTS

Both COSPEDTree and COSPEDTree-II are implemented in Python (version 2.7). Tree topologies are processed by the phylogenetic library Dendropy [20]. A desktop having Intel® Quad CoreIntel™ i5-3470 CPU, with 3.2 GHz processor and 8 GB RAM, is used to execute these methods.

A. Dataset

COSPEDTree-II is tested with the datasets like Marsupials (267 taxa and 158 input trees) [23], Placental Mammals (726 trees and 116 taxa) [24], Seabirds (121 taxa and 7 trees) [26], Temperate Herbaceous Papilionoid Legumes (THPL) (19 trees and 558 taxa) [27]. Work in [16] modified these datasets by removing duplicate taxon names and few infrequent taxa information¹. We have also experimented with Mammal dataset [12], [4] consisting of 12958 trees and 33 taxa². In addition, the dataset Cetartiodactyla (201 input trees and 299 taxa) [28] is also tested³.

B. Performance measures

Performance comparison between COSPEDTree-II and the reference approaches, employs the following measures:

¹Datasets are downloaded from the link <http://www.cs.utexas.edu/~phylo/software/superfine/submission/>.

²Downloaded from the link <http://www.supertriplets.univ-montp2.fr/>.

³Maintained in TreeBASE [29], and is downloaded from the link <https://treebase.org/treebase-web/search/study/summary.html?id=1271>.

- 1) **False positive distance** $FP(T, t_j)$: Number of internal branches present in the supertree T , but not in the input tree t_j ($1 \leq j \leq M$).
- 2) **False negative distance** $FN(T, t_j)$: Number of internal branches present in t_j but not in T .
- 3) **Robinson-Foulds distance** $RF(T, t_j)$: Defined as $FP(T, t_j) + FN(T, t_j)$.
- 4) **Maximum agreement subtree** $MAST(T, t_j)$: Let N_1 be the number of taxa contained in the maximum agreement subtree ($MAST$) common to T and t_j . Then, $MAST(T, t_j) = \frac{N_1}{|L(t_j)|}$. This measure is computed using Phylonet [30].

Above measures are accumulated for all the input trees t_j ($1 \leq j \leq M$), to be used as the final performance measures. Supertree producing lower values of the sum of FP , FN , and RF values is considered better. On the other hand, supertree having higher sum of $MAST$ score is considered superior.

C. Performance comparison

We have reported the results for the following two variations of COSPEDTree-II:

- 1) **COSPEDTree-II**: Produces supertree with possible multi-furcations.
- 2) **COSPEDTree-II + B**: Produces completely resolved binary supertrees, by applying the binary refinement suggested in section III-E.

Tables I to VI compare the performances of both of these variants, and with the reference approaches as well. Reference methods marked with a symbol ‘*’, could not be executed in all datasets, either due to the unavailability of corresponding source code, or due to their very high computational complexity. In such a case, we have used their results (both topological performance and running time) published in the existing studies [16]. The approaches MRP and superfine require PAUP* [31] to execute, which is a commercial tool and not available to us. So, these methods could not be tested in all datasets. Missing entries are indicated by ‘-’.

The methods RFS [15] and Supertriplet [4] produced errors in parsing few of the input datasets. Entries showing ‘ER’ indicate these errors. Supertrees generated by Supertriplet [4] could not be parsed by Phylonet [30]. So, we could not compute the $MAST$ scores for these trees. Finally, a symbol ‘F’ indicates that corresponding method could not produce a valid supertree for that dataset.

Results show that COSPEDTree-II produces better resolved supertrees than COSPEDTree, as indicated by lower FN , and mostly lower RF values for individual datasets. COSPEDTree-II also achieves higher $MAST$ scores for these datasets. COSPEDTree-II+B produces completely resolved binary supertrees. So, the number of FN branches reduces. However, as the input trees may not be fully resolved (may contain multi-furcating nodes), number of FP branches increases considerably. As COSPEDTree-II+B produces completely resolved supertrees, corresponding $MAST$ scores are much higher than COSPEDTree-II.

Comparison with reference approaches shows that only RFS [15] produces supertrees with consistently lower RF and higher $MAST$ scores than COSPEDTree-II. The method Superfine [16] performs better than COSPEDTree-II for the datasets Seabirds and THPL, while our methods perform slightly better (in terms of lower RF and higher $MAST$ score) for the Marsupials and Placental Mammals dataset. Superfine does not always generate strictly binary (completely resolved) supertrees (for example, in the THPL dataset), unlike COSPEDTree-II+B. Such a supertree exhibits much lower RF , but also much lower $MAST$ score (compared to COSPEDTree-II+B).

Matrix based methods like Minflip, SFIT, MMC, are outperformed by COSPEDTree-II. Veto approaches like SCM, PhysIC, produce supertrees with the lowest (mostly zero) FP branches, by not including any conflicting clades. In such a case, the number of FN branches becomes very high, and $MAST$ scores of these trees also become much lower. COSPEDTree-II also produces significantly better results than MRP PAUP for all the datasets except Cetartiodactyla. Subtree decomposition based approaches like thSPR, thTBR, produce slightly higher $MAST$ score values than COSPEDTree-II, since these methods directly synthesize input triplets, or in general, subtree topologies. Considering the measure RF , on the other hand, these methods are mostly outperformed by COSPEDTree-II.

D. Runtime Comparison

Tables I to VI express the running time of COSPEDTree-II and COSPEDTree-II+B for different datasets, in the formats (A+B) or (A+B+C), respectively, where:

- 1) A = Time to extract the couplet based measures from the input trees.
- 2) B = Time to process the couplets and cluster pairs, to produce a (possibly not binary) supertree.
- 3) C = Time required to refine the non-resolved supertree into a strict binary tree.

We observe that COSPEDTree-II incurs a significant fraction of its running time in the stages A and C. The stage A depends on the processing speed of the python based phylogenetic library Dendropy [20]. On the other hand, running time for the stage C depends both on the construction of $t'_{j|X}$ from individual t_j for all the multi-furcating nodes, and on the execution of thTBR [19]. Results show that COSPEDTree-II incurs much lower running time than COSPEDTree. Excluding the binary refinement stage, the running time is decreased by a factor from 2 (for the dataset Mammal [12], [4]) to 135 (for the dataset Cetartiodactyla [28]).

When the number of taxa is high (such as Marsupials [23], Cetartiodactyla [28]), COSPEDTree-II exhibits much lower running time than the triplet based methods [19], [4], due to its lower time complexity. For datasets with large number of trees, COSPEDTree-II incurs slightly higher running time than these methods, due to the time associated in extracting the couplet based measures.

V. CONCLUSION

We have proposed COSPEDTree-II, an improved couplet based supertree construction method (extending our earlier proposed method COSPEDTree). COSPEDTree-II produces supertrees with lower topological errors, and incurs much lower running time (compared to COSPEDTree). A binary refinement to generate a fully resolved supertree, is also suggested. Due to its high performance and much lower running time, COSPEDTree-II can be applied in large scale biological datasets.

EXECUTABLE

Executable and the results of COSPEDTree-II are provided in the link <http://www.facweb.iitkgp.ernet.in/~jay/phtree/cospedtree2/cospedtree2.html>.

ACKNOWLEDGEMENTS

The first author acknowledges Tata Consultancy Services (TCS) for providing the research scholarship.

REFERENCES

- [1] O. R. P. Bininda-Emonds, *An Introduction to Supertree Construction (and Partitioned Phylogenetic Analyses) with a View Toward the Distinction Between Gene Trees and Species Trees*, pp. 49–76. Springer, 2014.
- [2] A. D. Gordon, “Consensus supertrees: The synthesis of rooted trees containing overlapping sets of labeled leaves,” *Journal of classification*, vol. 3, pp. 335–348, 1986.
- [3] J. Dong, D. Fernández-Baca, and F. R. McMorris, “Constructing majority rule supertrees,” *Algorithms for Molecular Biology*, vol. 5, no. 2, pp. 1–16, 2010.
- [4] V. Ranwez, A. Criscuolo, and E. J. P. Douzery, “SuperTriplets: a triplet-based supertree approach to phylogenomics,” *Bioinformatics*, vol. 26, no. 12, pp. i115–i123, 2010.
- [5] R. Reaz, M. S. Bayzid, and M. S. Rahman, “Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach,” *Plos One*, vol. 9, no. 8:e104008, 2014.
- [6] S. Bhattacharyya and J. Mukherjee, “COSPEDTree: COuplet Supertree by Equivalence Partitioning of taxa set and DAG formation,” *IEEE/ACM Trans Comp Biol Bioinfo*, vol. 12, no. 3, pp. 590–603, 2015.
- [7] B. R. Baum, “Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees,” *Taxon*, vol. 41, pp. 3–10, 1992.
- [8] D. Chen, O. Eulenstein, D. Fernández-Baca, and M. Sanderson, “Minimum-Flip Supertrees: Complexity and Algorithms,” *IEEE/ACM Trans Comp Biol Bioinfo*, vol. 3, no. 2, pp. 165–173, 2006.
- [9] C. J. Creevey and J. O. McInerney, “Clann: investigating phylogenetic information through supertree analyses,” *Bioinformatics*, vol. 21, no. 3, pp. 390–392, 2005.
- [10] C. Semple and M. Steel, “A supertree method for rooted trees,” *Discrete Applied Mathematics*, vol. 105, pp. 147–158, 2000.
- [11] R. D. M. Page, “Modified Mincut Supertrees,” in *WABI*, pp. 537–551, Springer-Verlag, 2002.
- [12] V. Ranwez, V. Berry, A. Criscuolo, P. H. Fabre, S. Guillemot, C. Scornavacca, and E. J. P. Douzery, “PhySIC: A Veto Supertree Method with Desirable Properties,” *Syst Biol*, vol. 56, no. 5, pp. 798–817, 2007.
- [13] U. Roshan, B. Moret, T. Williams, and T. Warnow, “Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees,” vol. 3 of *Proc. 3rd Computational Systems Bioinformatics Conf. (CSB’04)*, pp. 98–109, IEEE, 2004.
- [14] D. R. Robinson and L. R. Foulds, “Comparison of phylogenetic trees,” *Mathematical Biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.
- [15] M. S. Bansal, J. G. Burleigh, O. Eulenstein, and D. Fernández-Baca, “Robinson-Foulds Supertrees,” *Algorithms for Molecular Biology*, vol. 5, no. 18, pp. 1–12, 2010.
- [16] M. S. Swenson, R. Suri, C. R. Linder, and T. Warnow, “SuperFine: Fast and Accurate Supertree Estimation,” *Syst Biol*, vol. 61, pp. 1–14, 2011.
- [17] D. T. Neves, T. Warnow, J. L. Sobral, and K. Pingali, “Parallelizing SuperFine,” in *Proc. 27th Ann. ACM Symp. on Applied Computing, SAC’12*, pp. 1361–1367, ACM, 2012.
- [18] S. Snir and S. Rao, “Quartets MaxCut: a divide and conquer quartets algorithm,” *IEEE/ACM Trans Comp Biol Bioinfo*, vol. 7, no. 4, pp. 704–718, 2010.
- [19] H. T. Lin, J. G. Burleigh, and O. Eulenstein, “Triplet supertree heuristics for the tree of life,” *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S8, 2009.
- [20] J. Sukumaran and M. T. Holder, “DendroPy: a Python library for phylogenetic computing,” *Bioinformatics*, vol. 26, no. 12, pp. 1569–1571, 2000.
- [21] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson, *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd ed., 2001.
- [22] L. Liu and L. Yu, “Estimating Species Trees from Unrooted Gene Trees,” *Syst. Biol.*, vol. 60, no. 5, pp. 661–667, 2011.
- [23] M. Cardillo, O. R. P. Bininda-Emonds, E. Boakes, and A. Purvis, “A species-level phylogenetic supertree of marsupials,” *J. Zool*, vol. 264, pp. 11–31, 2004.
- [24] R. Beck, O. R. P. Bininda-Emonds, M. Cardillo, F. R. Liu, and A. Purvis, “A higher-level MRP supertree of placental mammals,” *BMC Evol Biol*, vol. 6, no. 93, p. 14, 2006.
- [25] B. Holland, G. Conner, K. Huber, and V. Moulton, “Imputing Supertrees and Supernetworks from Quartets,” *Syst Biol*, vol. 56, no. 1, pp. 57–67, 2007.
- [26] M. Kennedy and R. Page, “Seabird supertrees: combining partial estimates of procariiform phylogeny,” *The Auk*, vol. 119, pp. 88–108, 2002.
- [27] M. Wojciechowski, M. Sanderson, K. Steele, and A. Liston, “Molecular phylogeny of the ‘temperate herbaceous tribes’ of papilionoid legumes: a supertree approach,” *Adv. Legume Syst*, vol. 9, pp. 277–298, 2000.
- [28] S. A. Price, O. R. P. Bininda-Emonds, and J. L. Gittleman, “A complete phylogeny of the whales, dolphins and even-toed hoofed mammals (Cetartiodactyla),” *Biological Reviews*, vol. 80, no. 3, pp. 445–473, 2005.
- [29] M. J. Sanderson, M. J. Donoghue, W. H. Piel, and T. Eriksson, “Tree-BASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life,” *American Journal of Botany*, vol. 81, no. 6, p. 183, 1994.
- [30] C. Than, D. Ruths, and L. Nakhleh, “PhyloNet: A Software Package for Analyzing and Reconstructing Reticulate Evolutionary Relationships,” *BMC Bioinformatics*, vol. 9, p. 322, 2008.
- [31] D. L. Swofford, “PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods),” 2003.